Electronic Prognostics

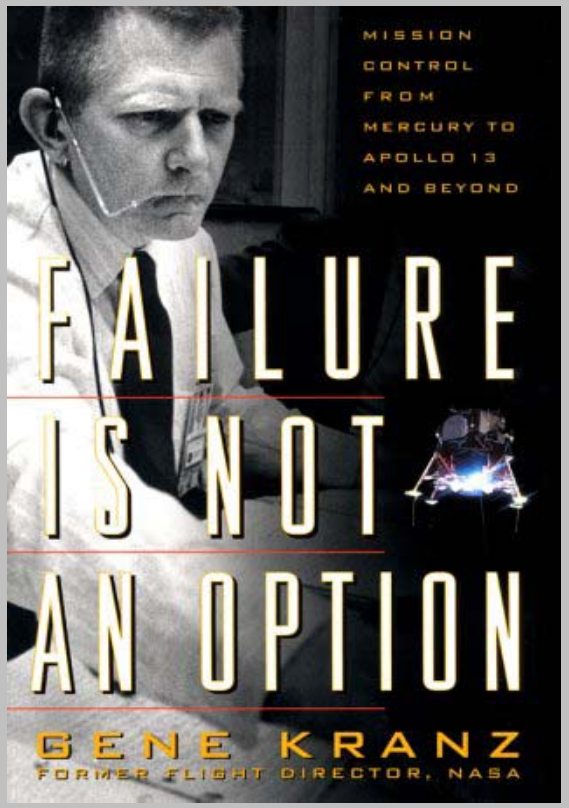
Continuous system telemetry coupled with real-time pattern recognition for enhanced reliability, availability and serviceability of electronic systems & networks

Kenny C. Gross
Sun Microsystems
San Diego Physical Sciences Research Center
1/23/06

Why is Availability Important?

- To Customers
  - Downtime very costly
    - Measured by lost sales, reduced productivity, damaged business reputation, diminished customer loyalty
- To Sun Microsystems
  - Means of differentiation
  - Key corporate initiative
  - Goal to provide continuous application access with predictable performance

In today's eCommerce-based computing model …



Mission director, Apollo 13

Motivation: Failures Are Costly

- Conventional approach for designing fault-tolerant and highly available systems: hardware redundancy
- Redundancy has several drawbacks:
  - Expensive:
    - replication of hardware components
    - multiplies power requirements
    - multiplies cooling requirements
    - increases complexity of heat removal
  - Complex: additional hardware/software required to implement failover, voting, state replication, checkpointing, etc.
  - Difficult to validate: realistic failure scenarios often difficult to mimic, making fault tolerance functionality one of the more undertested features of the system

Sun's Electronic Prognostics

- Application of advanced prognostic and "useful-life-remaining" modeling in support of Prognostics and Health Management (PHM) for electronic components, system boards, CPUs, networks, digital component elements and digital devices.
- With digital electronic boards playing a key role in the operation of future military electronic systems and subsystems, it is important that the user be able to accurately diagnose faults and predict failures and life remaining of these components.

Continuous System Telemetry Harness

"Soft" Variables

System Performance Variables from *kstat*

"Canary" Variables

Distributed Synthetic Transaction Generators (user wait times, monitored 24x7)

Physical Variables

Distributed internal temperatures, current & voltage time series

"Black Box" Recorder

Circular File Structure.

Retains high sampling rate signals 72 hrs; lower sampling rate signals 30 days

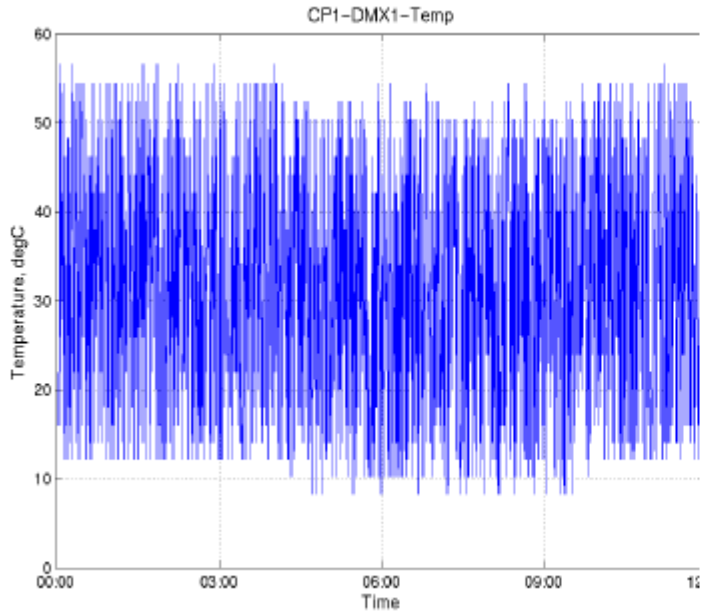
Consumer Processes for Telemetry Signals

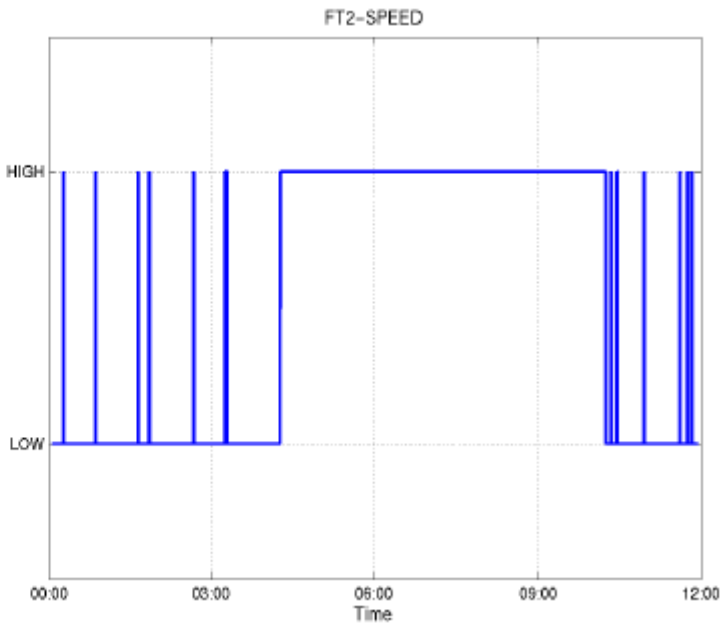
Proactive:	Predictive Failure Annunciation
Reactive:	Faster, more accurate root cause analysis
Self Healing:	Software Aging and Rejuvenation
NTF Reduction:	Captured signatures help reveal the mechanisms responsible for No-Trouble-Found (NTF) events

Failure Mechanisms Detectable Through CSTH + MSET

- Delamination of bonded IC components
- Departure from coplanarity in stacked components
- Solder joint cracking
- Interconnect degradation for optical and mechanical sockets
- Electrostatic discharge (ESD) phenomena
- Propagation of vibrational resonances
- Wearout of components in electromechanical systems (disk drives, fan trays, turbo blowers)
- Leakage of electrolyte in electrolytic capacitors
- Capacitors installed by vendor with reversed polarity
- Spontaneous fan oscillations from sensor degradation, producing excessive thermal cycling and accelerated aging problems
- Thermal issues because user forgot to change air filters
- Contaminated molding compounds in IC chips and memory modules
- Sensor failures in general (Sun Fire 15K has > 1000 sensors; simple arithmetic with MTBFs indicate that many of those sensor will fail over a 5 yr projected life of a server, yet w/o telemetry there is no way to tell a sensor is degrading or “stuck”)
- Software aging problems (memory leaks have an average of 56 days for “time to customer relief”)
- Power supplies with all types of AC/DC and DC/DC PSU degradation modes
- Bit error rates in interconnects and FCAL loops
- Fan tray failures that go undetected by expensive RPM sensors because the wind in the machine keeps the blades turning

Early Telemetry Success on SunFire F15K

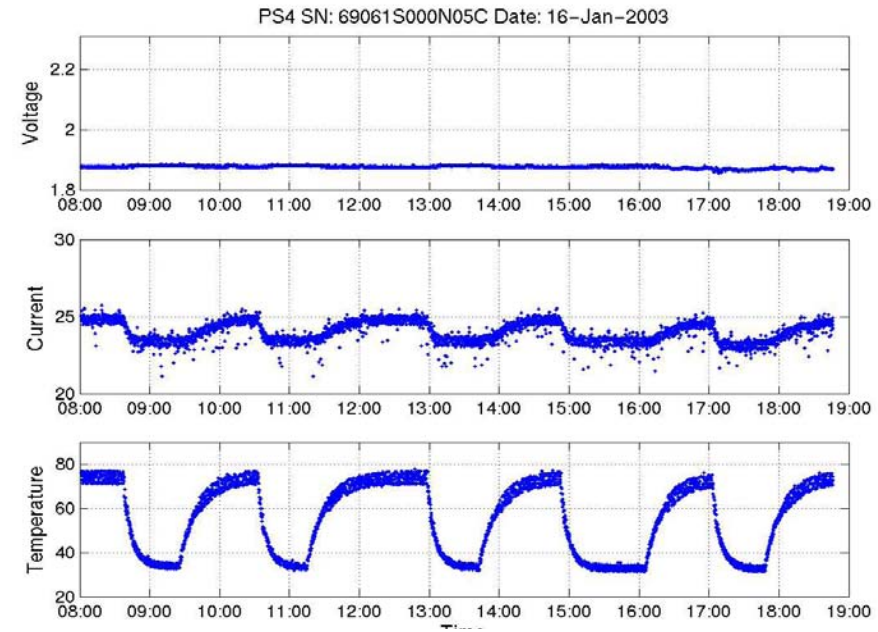


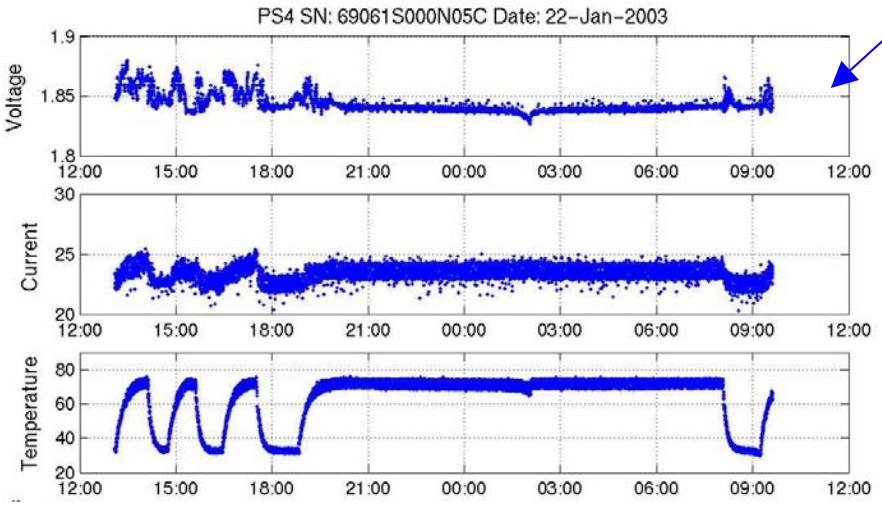


SMS reporting processor temp incorrectly (note wild swings 12-50°C)

Fans continuously cycle between low and high (1 of 8 fans shown)

CSTH for NTF Mitagation





Ongoing thermal cycling experiments with NTF power supplies from E10Ks

Upper plots show flat voltage during temperature cycling with undegraded power supply.

Lower plots show voltage fluctuations from degrading power supply.

Voltage fluctuations from degrading power supplies are causing the system boards to throw out a number of failure messages, including:

- DTAG failures
- DTAG parity errors
- Ecache Failures
- Coherent processor errors
- UPA fatal error
- UPA parity error

Advanced Pattern Recognition Tools for Ultrahigh-Reliability Surveillance

Sequential Probability Ratio Test (SPRT)

For Stationary Time Series

- Advanced pattern recognition technique for high sensitivity, high reliability sensor and equipment operability surveillance.
- Developers proved in refereed journals that the SPRT provides the earliest mathematically possible annunciation of a subtle fault in noisy process variables.

For Dynamic Time Series

Multivariate State Estimation Technique (MSET)

- Online model-based fault detection and identification.
- MSET predicts what each process should be on the basis of learned correlations among all process variables.
- MSET incorporates the SPRT to monitor the residuals between the actual observations and the estimates MSET predicts on the basis of the correlated variables.

AVOIDANCE OF FALSE ALARMS

MSET was developed for applications in which there may be enormous penalties associated with false alarms. Two examples:

Commercial Nuclear Industry:

False alarms result in plant shutdowns, which cost $1M per day. EPRI studies have shown that 25% of nuclear plant trips are false alarms from degrading sensors (many sensors have much shorter MTBFs than the assets they are monitoring).

Result: After a rigorous 2 year evaluation, the US NRC formally approved (2/16/00) the use of MSET for continuous calibration validation of all safety-critical and life-critical sensors in all US nuclear plants.

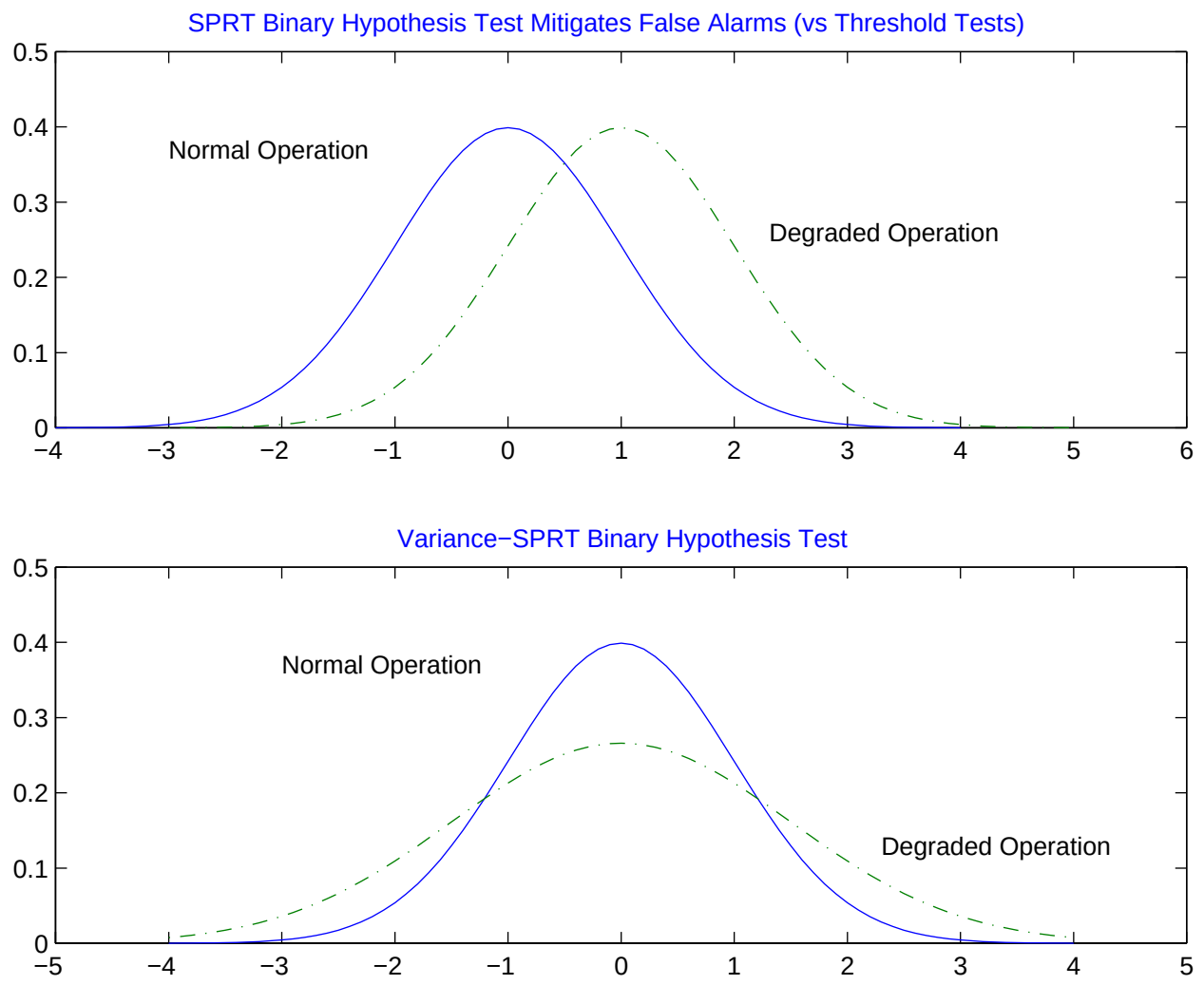
NASA Space Shuttle:

Although many shuttle countdowns have been aborted prior to liftoff, one shuttle mission was aborted 2 seconds after liftoff in 1985.

For an abort-after-liftoff, the shuttle is dumped into the ocean. The cost of this incident was $50M to NASA.

Postmortem analysis showed that 2 redundant vibration sensors had malfunctioned simultaneously, falsely indicating an anomalous surge in engine vibration. There was no engine problem.

Result: NASA awarded $700K in contracts since to adapt MSET for continuous surveillance of all sensors and components on space shuttle main launch vehicles.

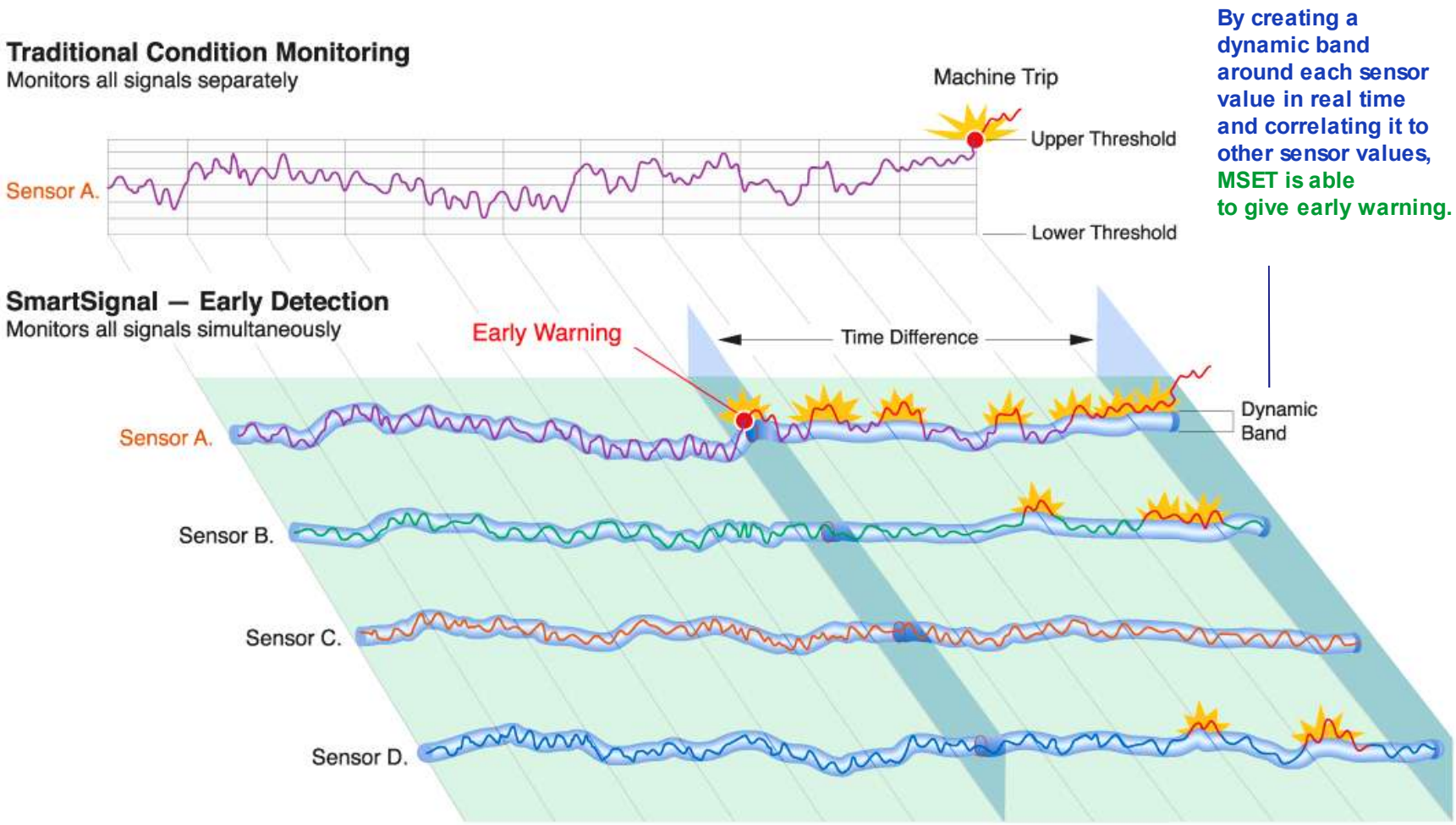

SPRT Binary Hypothesis Test Mitigates False Alarms (vs Threshold Tests)
Normal Operation
Degraded Operation
0.5
0.4
0.3
0.2
0.1
0
−4
−3
−2
−1
0
1
2
3
4
5
6
Variance−SPRT Binary Hypothesis Test
Normal Operation
Degraded Operation
0.5
0.4
0.3
0.2
0.1
0
−5
−4
−3
−2
−1
0
1
2
3
4
5

What is MSET?

Multivariate State Estimation Technique

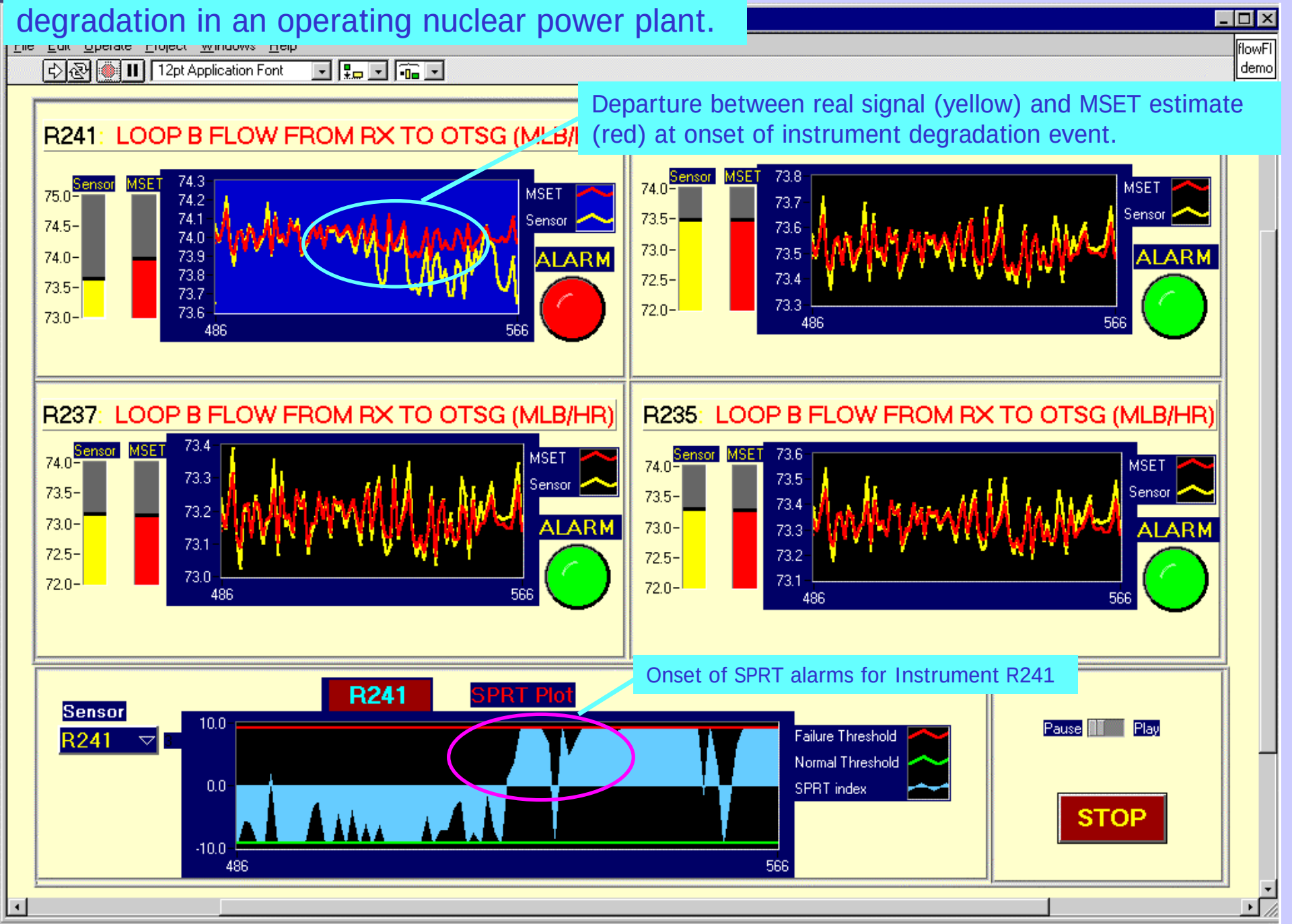
- Advanced pattern recognition system developed for 24x7 predictive fault monitoring in complex engineering systems like avionics and nuclear reactors
  - Continuous signal and sensor operability validation
  - Incipient fault annunciation on all monitored components
  - Extremely low probability of false alarms
- Award winning incipient fault surveillance system developed by Argonne National Laboratory
  - Capabilities surpass conventional pattern recognition approaches, including neural networks, in sensitivity, reliability, and computational efficiency

CSTH + MSET Pattern Recognition: Early Detection



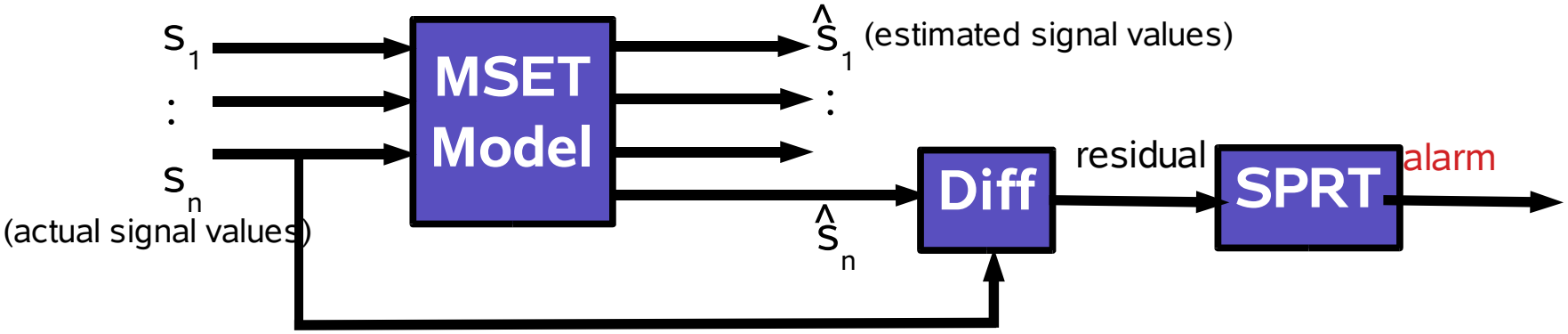
By creating a dynamic band around each sensor value in real time and correlating it to other sensor values, MSET is able to give early warning.

Legacy Viewgraph: MSET detects instrumentation degradation in an operating nuclear power plant.



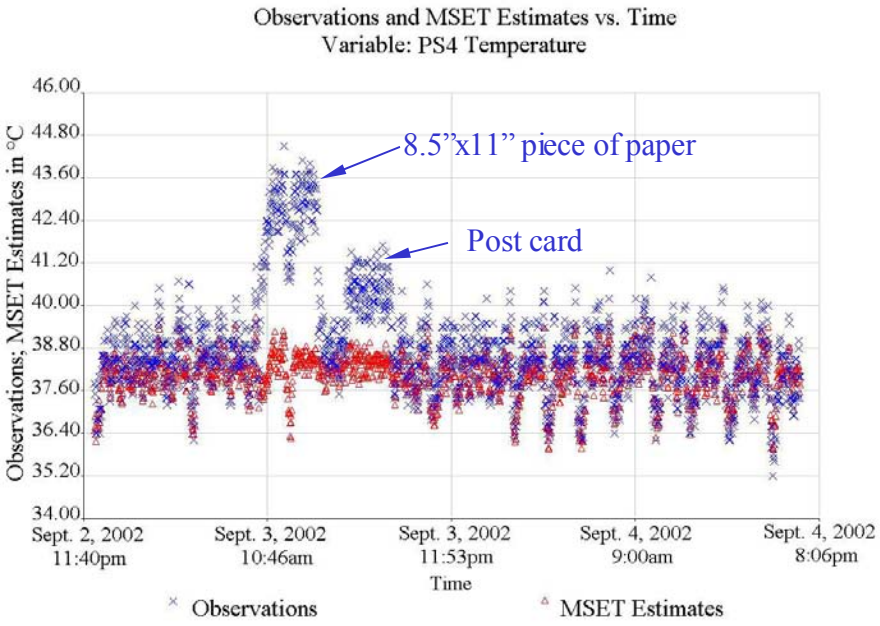
MSET In Action

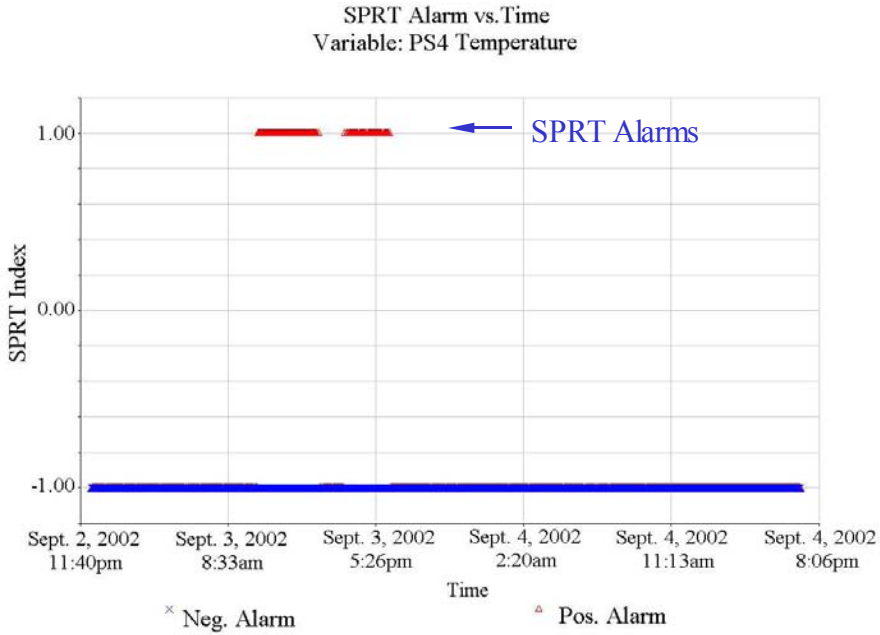
- Train on signal data collected under expected operational profile of system
- MSET pattern recognition learns correlations among signals, creating model of the system
- In monitoring mode, actual signals compared with estimates produced from model:



Thermal Anomalies Lead to Electronic System Failures

- High temperature protection thresholds quite high (85 – 110°C)
- Many customer service calls originate from thermal problems that have much lower temperatures, but lead to long-term reliability issues:
  - Failing to change air filters
  - Running cables in raised-floor cool-air channel
  - Scrap paper sucked onto bottom air inlet grill
  - Inadvertently configuring hot-air exhaust from one machine to cold-air inlet of another
- MSET detects thermal anomalies with high sensitivity and minimal false alarm probability





MSET Detects Coolant Air Flow Perturbations in Enterprise Servers

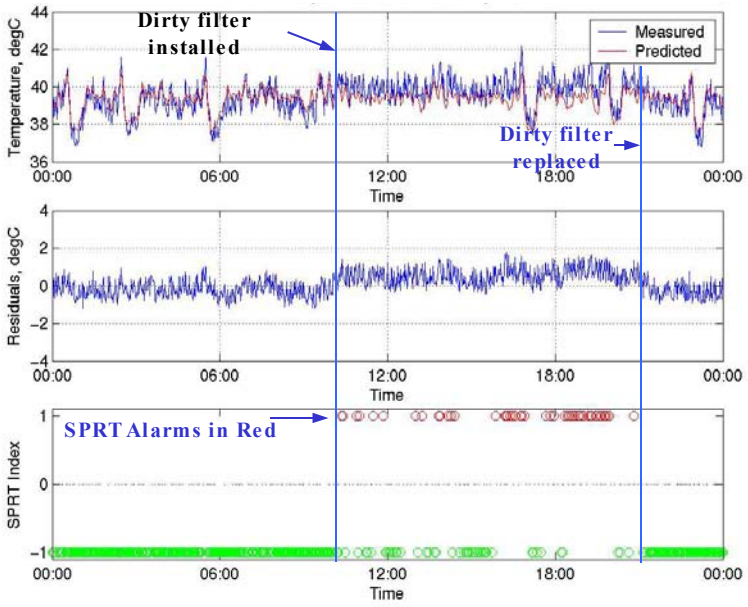
Occasional cause of service problems with high end servers: piece of paper falls from wall, notebook, etc. Works its way to bottom air inlet for server. Temps are not high enough to trip threshold; but over long term, can lead to accelerated reliability issues.

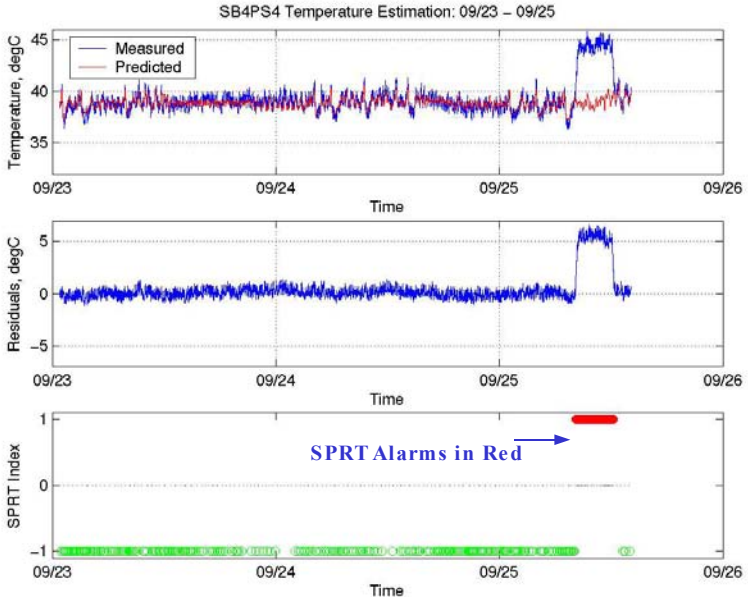
Experiments conducted with fully loaded E10K. MSET monitors dozens of performance variables. Piece of paper put on bottom air inlet.

Immediate SPRT alarms observed.

2nd experiment conducted with 3x5 post card.

SB4PS4 Temperature: Actual and MSET Estimate





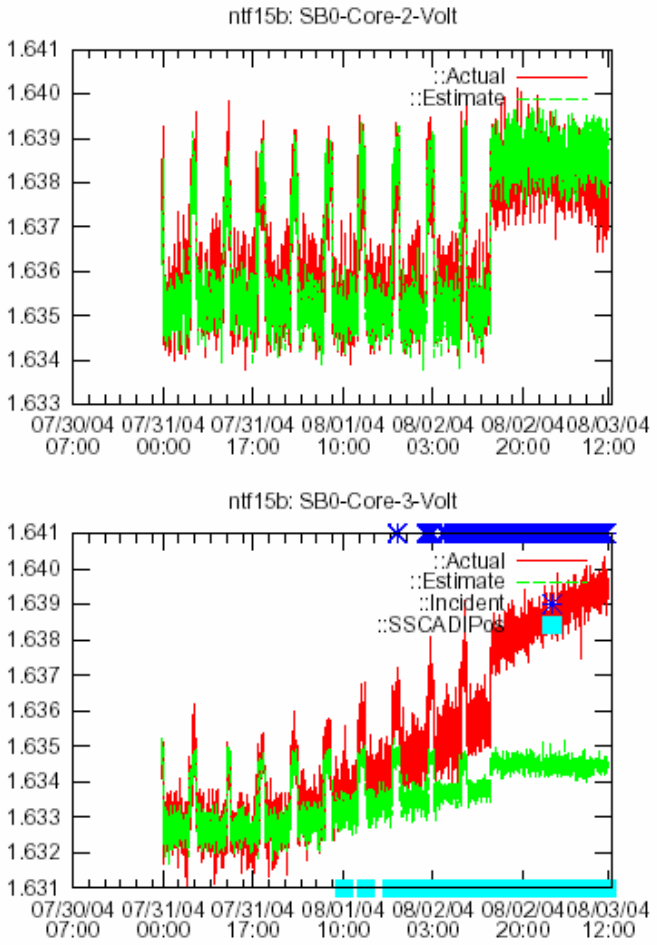
DATE: September 23 – 26 2002

MSET Detects Fouled Air Filters in Enterprise Servers

MSET Detects Degraded/Failed Fans

(Eliminates Need for Hall-Effect RPM Sensors)

MSET Detects Drifting Power Supply Voltages



Can detect anomalies in nominally stationary signals.

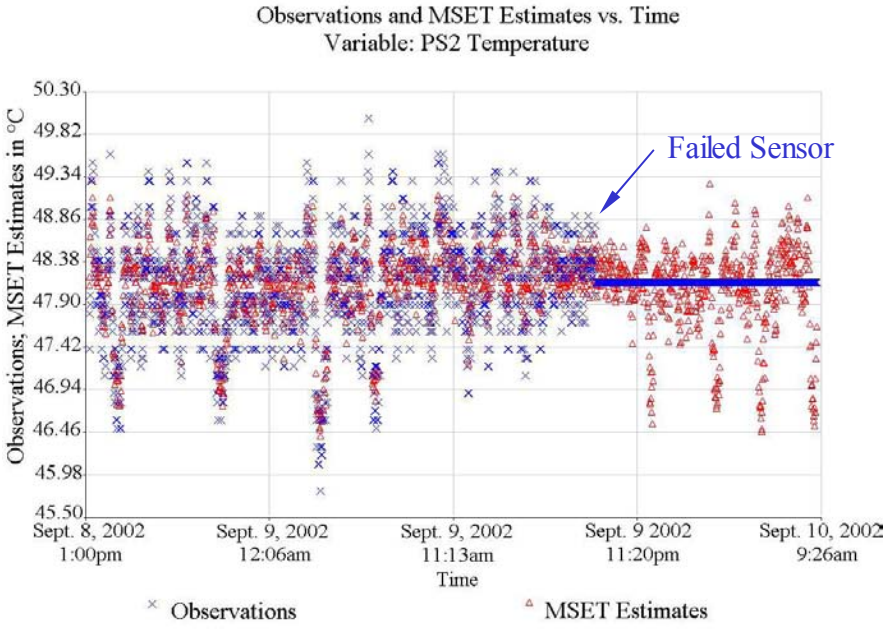
MSET automatically takes into account load effects by monitoring correlated signals.

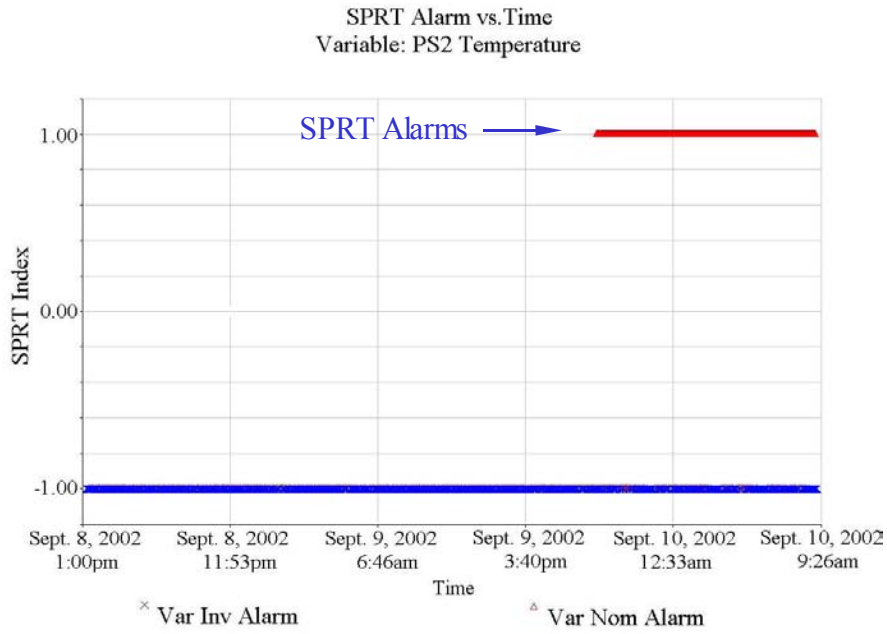
Inferential Sensing

Sun's high-end servers contain hundreds of physical sensors (distributed board, module, and ASIC temperature sensors, voltages, and currents) that protect the system by detecting when a parameter is out of bounds, and then shutting down a component, system board, domain, or entire system.

When a sensor failure is detected, a pattern recognition module swaps out the degraded sensor signal, and swaps in an "analytical estimate" of the physical variable. The analytical estimate is supplied by the pattern recognition algorithm and is called an "inferential sensor". This analytical estimate can be used indefinitely, or until the Field Replaceable Unit (FRU) containing the failed sensor needs to be replaced for other reasons.

Inferential Sensors via MSET





Physical sensors can fail. In many cases, the physical sensors have a shorter MTBF than the assets the sensors are supposed to protect.

With MSET, if a physical sensor fails or degrades in service, MSET can mask the sensor signal and

swap in the MSET estimate (red variable in figure).

Immediate SPRT alarms observed.

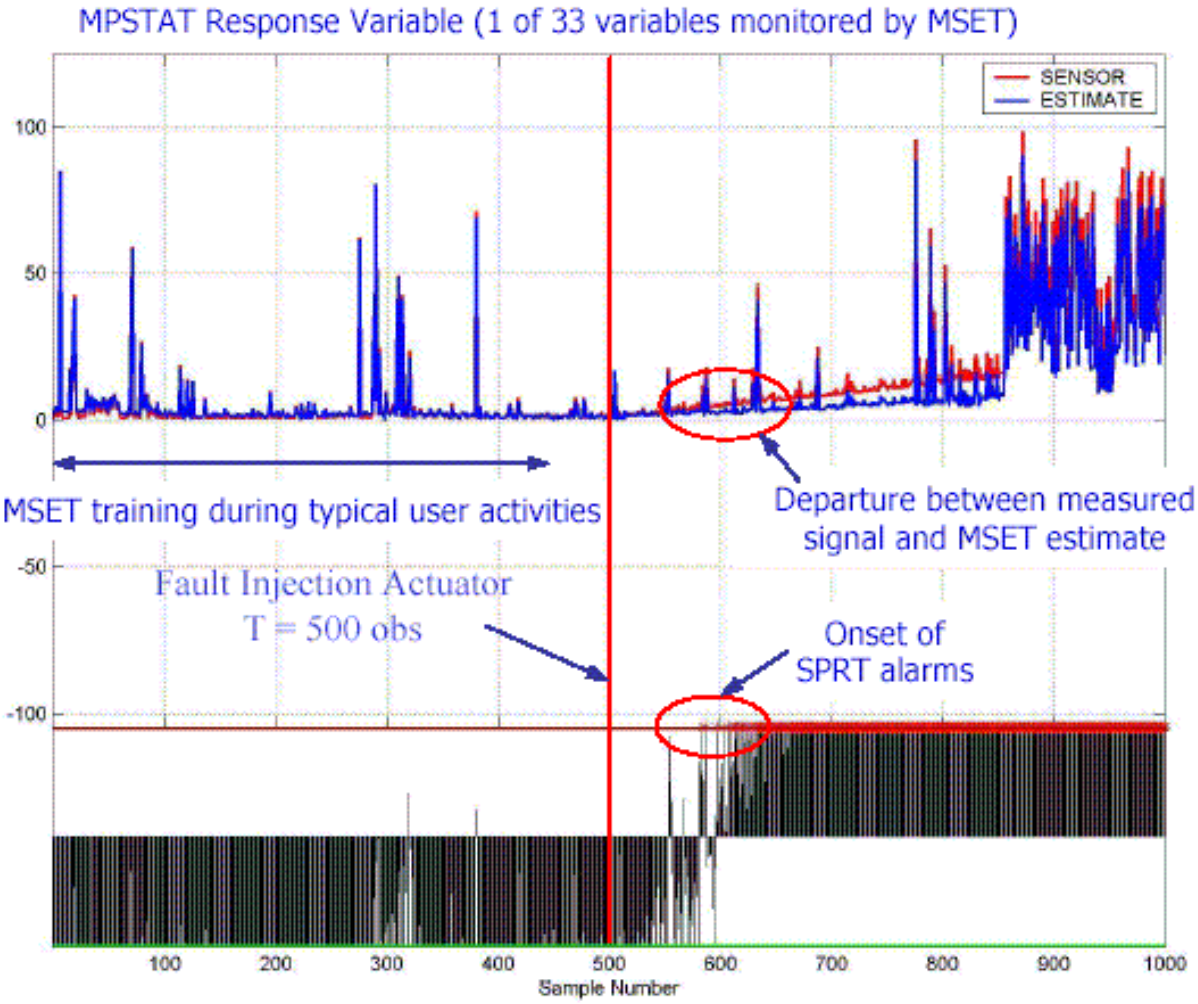
Realtime Sensor Validation: Benefits

- All control actuator functions now use fully validated signals
- For many (perhaps most) industrial systems, including Sun high-end servers, the sensors often have shorter MTBFs than the assets they are supposed to protect
- MSET has a unique capability, called inferential sensing, to detect the onset of sensor degradation and swap in a highly accurate analytical estimate. Sensor replacement can be postponed until the next scheduled outage.

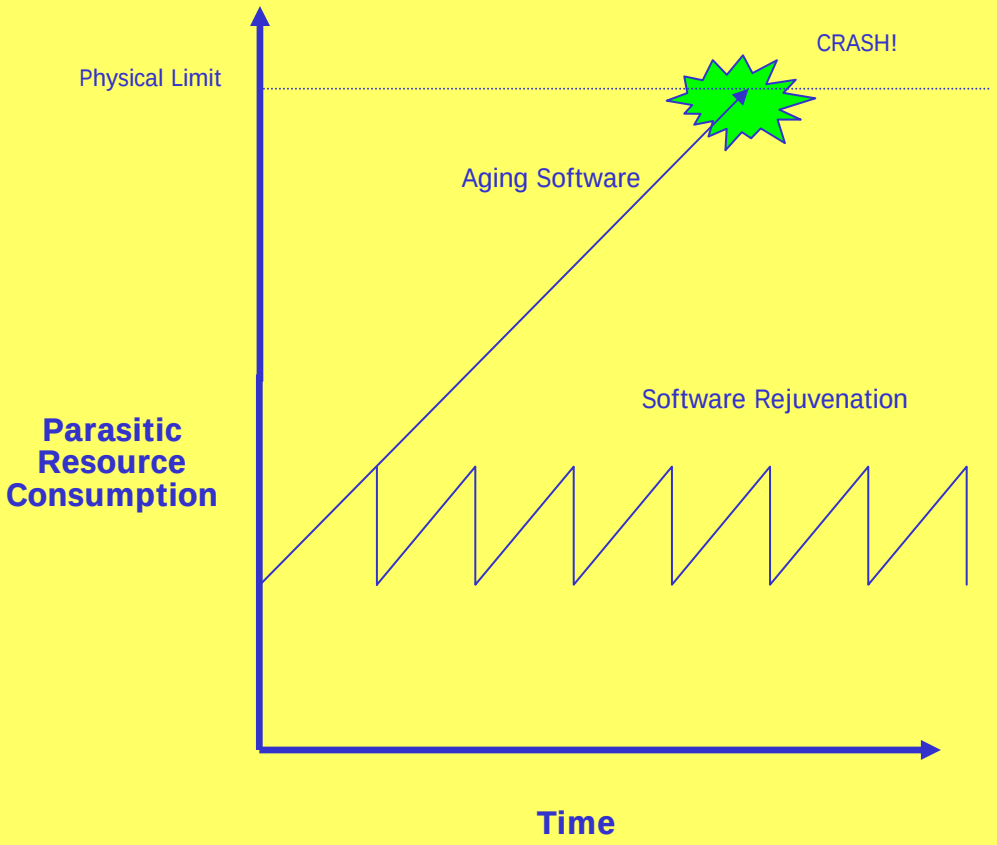
Software Aging and Rejuvenation

- Software aging: long-running applications degrade over time:
  - Memory leaks, unreleased file locks, accumulation of unterminated threads, data corruption/roundoff accrual, file space fragmentation, shared memory pool latching, thread stack bloating and overrun
- Software rejuvenation: periodically “cleanse” internal system state to prevent software aging effects:
  - Flush stale locks, reinitialize application components, preemptive rollback, defragment disk, therapeutic reboot (primarly Windows platform)

MSET Detects Onset of Software Aging

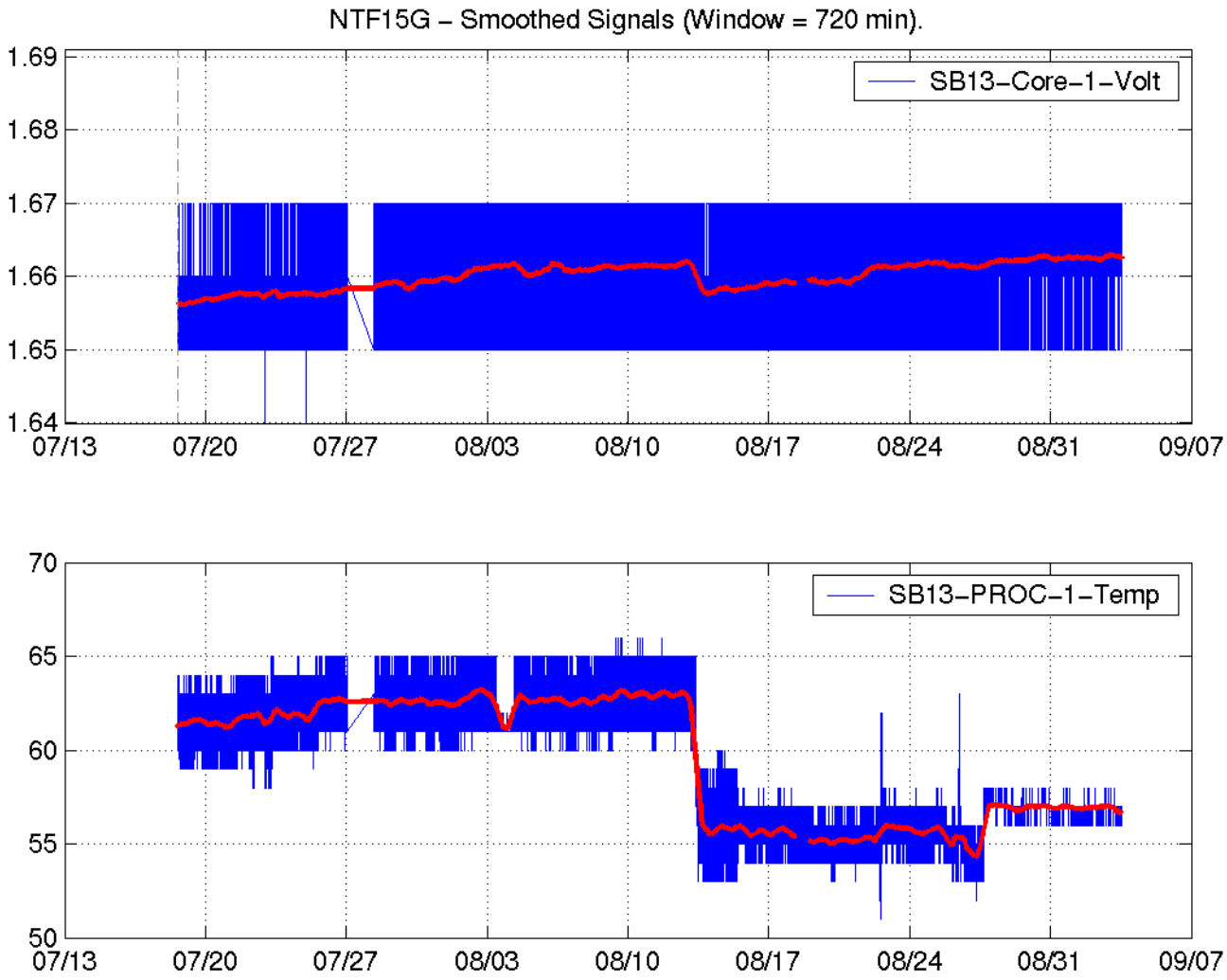


Software Rejuvenation Reduces Downtime



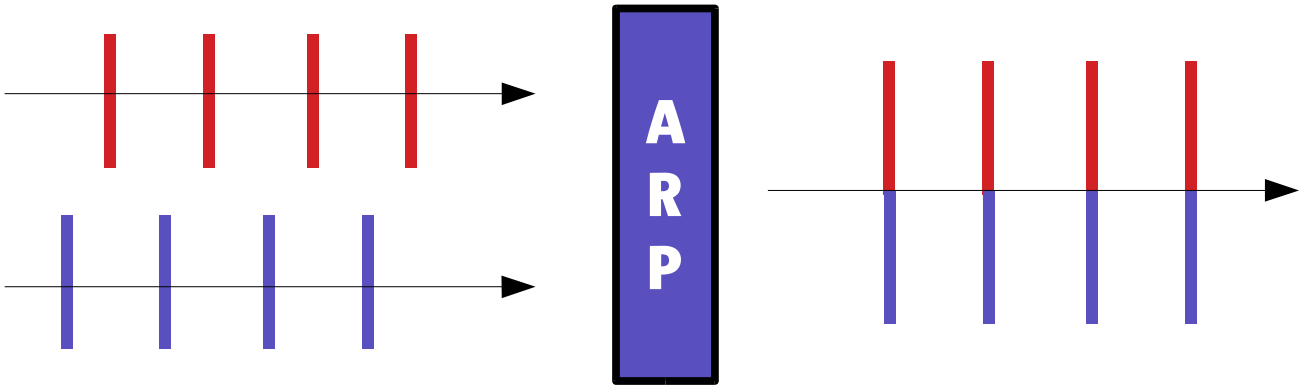
Challenge: Quantized Signals

Blue signals show the raw signals reported from 8-bit A/D chips used in most computing systems. Upper plot is a typical voltage, lower plot is a typical temperature. The red signal shows the actual value of the variable being monitored. For future systems Sun recommends 12-bit A/D (minimum). For legacy systems with 8-bit A/D, Sun has a proprietary "Moving Histogram" method to attain high-accuracy prognostics from low resolution sensors.



Unifying Disparate Telemetry Streams

- Groups of sensors may be sampled by multiple threads (e.g., domain-side monitors of CPU utilization and SP-side monitors of temperature)
- Each Detector requires unified telemetry stream from a Collector
- Analytical Resampling Process (ARP) preprocesses data to line up disparate telemetry streams



TIRF: Telemetric Impulsional Response Fingerprint

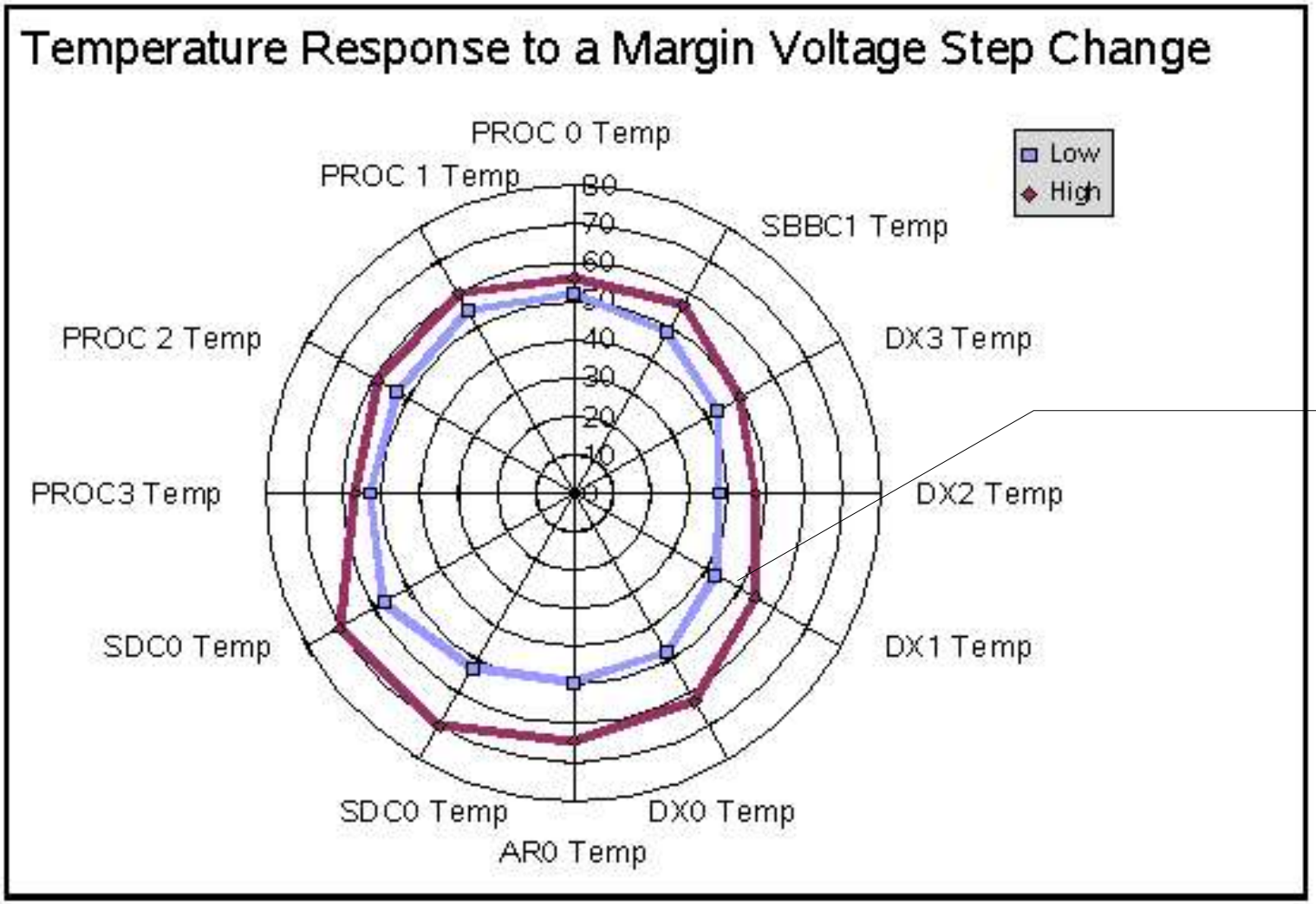
Technique:

- TIRF is a response of a component (FRU) to a controlled perturbation

- Examples:
  - Response of StarCat system board CPU temperatures to power supply step voltage change
  - Response of a set of software or QoS variables to a step change in computational load

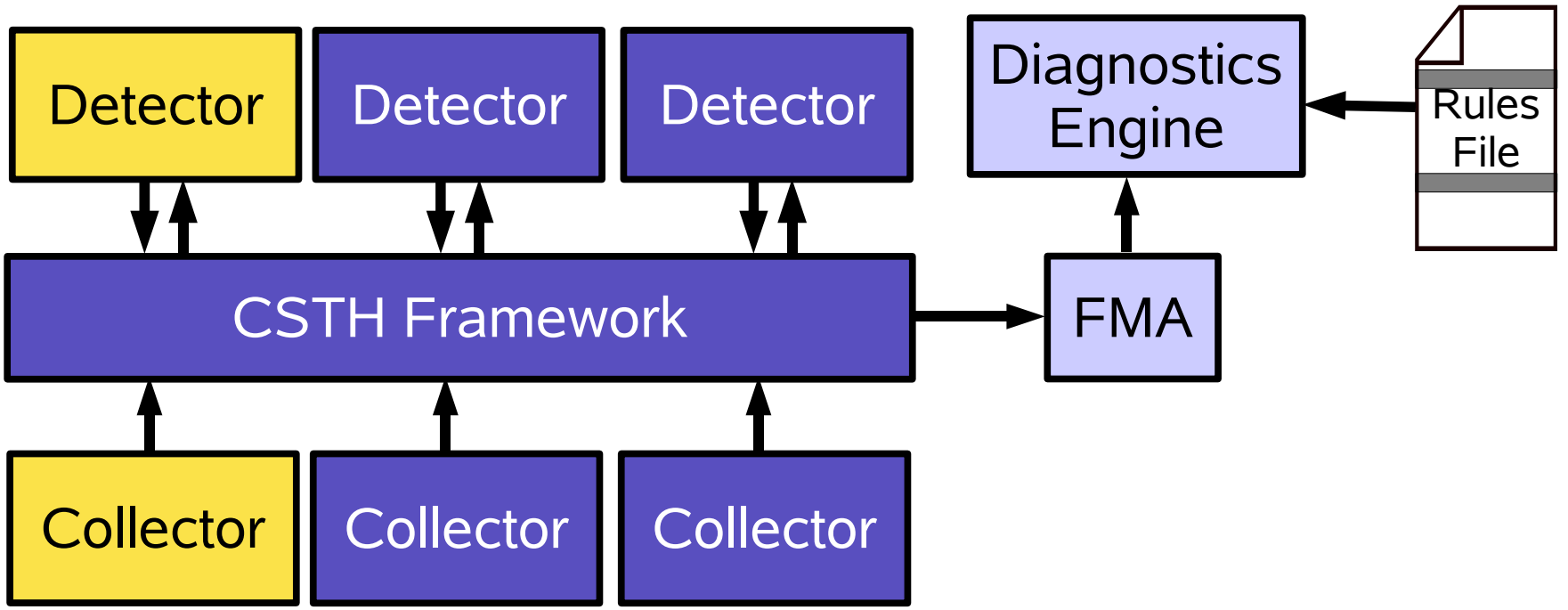
- Purpose:
  - Active probing for detection of dormant faults
  - Universal feature extractor for black box modeling
  - TIRFs couple well with pattern recognition for mitigation of No-Trouble-Found (NTF) events

TIRF: Spider Plot Representation of Response



Response is used as features for pattern recognition and classification in fault detection applications

Telemetry Capabilities Upgraded on Demand



- When new monitoring functionality required, new Collector, Detector, and rules set can be plugged into telemetry framework

CSTH Lab Console Displays For Coordination Of Dynamical System Characterization Experiments

Summary: Electronic Prognostics

Continuous System Telemetry Harness

- Provides "black box flight recorder" for components, systems, peripherals, and associated networks
- System monitoring and analysis methodology
  - **Proactive fault monitoring** (detect incipient failures)
  - Faster, more accurate Root Cause Analysis (RCA)
  - Mitigate No Trouble Found (NTF) parts returns
  - Uses sophisticated pattern recognition techniques (called MSET)
    - Provides real time "Intelligent Agent" functionality
    - Used for failure prediction, resource management, condition-based maintenance
    - Enhances reliability, availability, and serviceability (RAS)
  - Future servers and networks will use this technique for closed-loop autonomic control

Bibliography

Sun Microsystems' Proactive Fault Monitoring Innovations

- "Electronic Prognostics Through Continuous System Telemetry," K. C. Gross, K. W. Whisnant and A. Urmanov, ***Proc. 60th Meeting of the Society for Machinery Failure Prevention Technology***, Virginia Beach, VA (April 2006). [Accepted for publication]

- "Incipient Fault Detection in Storage Systems using On-Line Pattern Recognition" K. Vaidyanathan, K. C. Gross and R. Dhanekula, ***Proc. 60th Meeting of the Society for Machinery Failure Prevention Technology***, Virginia Beach, VA (April 2006). [Accepted for publication]

- "Electronic Prognostics Techniques for Mission Critical Electronic Components and Subsystems," K. C. Gross, K. W. Whisnant and A. M. Urmanov, Proc. ***2006 Components for Military and Space Electronics Symposium,*** Los Angeles, CA, (Feb 2006). [Accepted for publication]

- "Monte Carlo Simulation of Telemetric Signals for Enhanced Proactive Fault Monitoring of Computer Servers," K. Vaidyanathan and K. C. Gross, ***Proc. 2005 Simulation Multiconference,*** Philadelphia, PA (July 2005).

- "Continuous System Telemetry Harness: Field Failure Prediction and Data Collection Methodology," K. Neuman, ***Proc. 2005 ASME Integrated System Packaging Conference (InterPACK05)***, San Francisco, CA (July 2005).

- "Proactive Fault Monitoring in Enterprise Servers," K. Whisnant, K. C. Gross and N. Lingurovska, ***Proc. 2005 IEEE Intn'l Multiconference in Computer Science & Computer Eng.***, Las Vegas, NV (June 2005).

- "Dynamic System Characterization of Enterprise Servers via Nonparametric Identification," E. Schuster and K. C. Gross, ***Proc. 2005 American Control Conf.***, Portland, OR (June 2005).

- "Spectral Decomposition and Reconstruction of Telemetry Signals from Enterprise Computing Systems," K. C. Gross and E. Schuster, ***Proc. 2005 IEEE Intn'l Multiconference in Computer Science & Computer Eng.***, Las Vegas, NV (June 2005).

- "Monte Carlo Simulation for Optimized Sensitivity of Online Proactive Fault Monitoring Schemes," K. Vaidyanathan and K. C. Gross, ***Proc. 2005 Intn'l Conf. on Computer Design***, Las Vegas, NV (June 2005).

- "Failure Avoidance in Computer Systems," A. Urmanov and K. C. Gross, ***Proc. 59th Meeting of the Society for Machinery Failure Prevention Technology,*** Virginia Beach, VA (Apr 18-21, 2005).

- "Proactive Fault Monitoring Using Advanced Pattern Recognition Algorithms," N. Lingurovska, BS Thesis, Kettering University (Oct 2004).

- "Proactive Detection of Software Anomalies through MSET," K. Vaidyanathan and K. C. Gross, ***Proc. IEEE Workshop on Predictive Software Models (PSM-2004)***, Chicago (Sept 17-19, 2004).

- "Multi-Frequency Sinusoidal Perturbation Method for Dynamic Characterization of Multi-Processor Computer Servers," E. Schuster and K. C. Gross, ***SunLabs Technical Report TR-2004-130***, Sun Microsystems (June 2004).

- "A New Sensor Validation Technique for the Enhanced RAS of High End Servers," A. Urmanov, B. Guenin, K. C. Gross, and A. Gribok, ***2004 Intn'l Conf. on Machine Learning; Models, Technologies and Applications (MLMTA'04)***, Las Vegas, NV (June 21 - 24, 2004).

- "A New Framework for Analyzing Complex Networks of Entities," A. Urmanov, A. Bougaev, K. C. Gross, and A. Gribok, ***2004 Intn'l Conf. on Machine Learning, Models, Technologies and Applications (MLMTA'04)***, Las Vegas, NV (June 21 - 24, 2004).

- "Neutron Measurements for Improved Soft Error Discrimination in Computer Systems," A. Urmanov, K. C. Gross and A. Bougaev, ***Transactions of the Amer. Nuclear Soc.,*** Pittsburgh, PA (June 13-17, 2004).

- "Improved Methods for Early Fault Detection in Enterprise Computing Servers Using SAS Tools," K. C. Gross and K. Mishra, ***2004 SAS Users Group International*** *(SUGI 29)*, Montreal, Canada. (May 9 – 12, 2004).

- "Dynamic Stimulation Tool for Improved Performance Modeling and Resource Provisioning of Enterprise Servers," K. Mishra and K. C. Gross, ***Proc. 14th IEEE Intn'l. Symp. on Software Reliability Eng. (ISSRE'03),*** Denver, CO (Nov. 2003).

- "MSET Performance Optimization for Detection of Software Aging," K. Vaidyanathan and K. C. Gross, ***Proc. 14th IEEE Intn'l. Symp. on Software Reliability Eng. (ISSRE'03),*** Denver, CO (Nov. 2003).

- "Frequency-Domain Pattern Recognition for Dynamic System Characterization of Enterprise Servers," K. C. Gross and K. Mishra, ***Proc. 2003 Intn'l Conf. on Artificial Intelligence (IC-AI'03),*** Las Vegas, NV (June 23-26, 2003).

- "Remote Measurement and Analysis of Executing Software Systems," A. Porter, S. McMaster, A. Urmanov, L. G. Votta and K. C. Gross, ***Proc. 2003 Intn'l Conf. on Software Engineering (ICSE03),*** Portland OR (May 9, 2003).

- "Proactive System Maintenance Using Software Telemetry," K. C. Gross, S. McMaster, A. Porter, A. Urmanov, and L. G. Votta, in A. Osslo and A. Porter, editors, ***Proc. Remote Analysis and Measurement of Software Systems*** (May 2003)

- "Spectral Decomposition of Performance Variables for Dynamic System Characterization of Web Servers," K. C. Gross, W. Lu, and K. Mishra, ***Proc. 2003 SAS Users Group International (SUGI 28)***, Seattle, Wa. (Mar 30 - Apr 2, 2003).

- "Proactive Detection of Software Aging Mechanisms in Performance-Critical Computers," K. C. Gross, V. Bhardwaj, and R. L. Bickford, 27th Annual IEEE/NASA Software Engineering Symposium, Greenbelt, MD (Dec 4-6, 2002).

- "Time-Series Investigation of Anomalous CRC Error Patterns in Fibre Channel Arbitrated Loops," K. C. Gross, W. Lu, and D. Huang, ***Proc. 2002 IEEE Int'l Conf. on Machine Learning and Applications (ICMLA),*** Las Vegas, NV (June 2002).

- "Early Detection of Signal and Process Anomalies in Enterprise Computing Systems," K. C. Gross and W. Lu, ***Proc. 2002 IEEE Int'l Conf. on Machine Learning and Applications (ICMLA),*** Las Vegas, NV (June 2002).

- "Advanced Pattern Recognition for Detection of Complex Software Aging Phenomena in Online Transaction Processing Servers," Karen J. Cassidy, Kenny C. Gross, and Amir Malekpour, ***Proc. Intnl. Performance and Dependability Symposium,*** Washington, DC, (June 23rd - 26th, 2002).

- "Software Reliability and System Availability at Sun," K. C. Gross, ***Proc. IEEE 11th Intn'l Symp. On Software Reliability Eng.***, San Jose, CA (Oct 2000).